# Foreebank: Syntactic Analysis of Customer Support Forums

Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn, Joseph Le Roux

# Foreebank: Syntactic Analysis of Customer Support Forums

**Rasoul Kaljahi**[1], **Jennifer Foster**[1] **Johann Roturier**[2]
**Corentin Ribeyre**[3], **Teresa Lynn**[1], **Joseph Le Roux**[4]
[1]ADAPT Centre, School of Computing, Dublin City University, Ireland
{rkaljahi, jfoster, tlynn}@computing.dcu.ie
[2]Symantec Research Labs, Dublin, Ireland
johann_roturier@symantec.com
[3]Alpage, INRIA, Univ Paris Diderot, Sorbonne Paris Cité, France
corentin.ribeyre@inria.fr
[4]Université Paris Nord, France
joseph.le.roux@gmail.com

## Abstract

We present a new treebank of English and French technical forum content which has been annotated for grammatical errors and phrase structure. This double annotation allows us to empirically measure the effect of errors on parsing performance. While it is slightly easier to parse the corrected versions of the forum sentences, the errors are not the main factor in making this kind of text hard to parse.

## 1 Introduction

The last five years has seen a considerable amount of research carried out on web and social media text parsing, with new treebanks being created (Foster et al., 2011; Seddah et al., 2012; Mott et al., 2012; Kong et al., 2014), and new parsing systems developed (Petrov and McDonald, 2012; Kong et al., 2014). In this paper we explore a particular source of user-generated text, namely, posts from technical support forums, which are a popular means for customers to resolve their queries about a product. An accurate parser for this kind of text can be used to inform forum-level question-answering, machine translation and quality estimation of machine translation.

We create a 2000-sentence treebank called *Foreebank* which contains sentences from the Symantec Norton English and French technical support forums.[1] The phrase structure of the sentences is annotated *and* any grammatical errors are marked in the trees. Marking the grammatical errors during the process of syntactic annotation allows us to precisely measure the amount of grammatical noise in this kind of text and also to determine its effect on parsing.

Foster (2010) explored the effect of spelling errors on parsing performance of conversational forum text. We extend this study to include grammatical errors, focusing on more technical content. Foster et al. (2008) explored the effect of artificially generated grammatical errors on Wall Street Journal parsing. We concentrate on forum text rather than newspaper text, and, crucially, examine the effect of *real* grammatical errors. We find that the level of grammatical noise is lower than expected, with capitalisation and punctuation errors being the most frequent. While correcting all the errors does result in a performance increase of 1.5% for English and 0.8% for French, the major challenge in parsing these sentences seems not to be "bad language" (Eisenstein, 2013) per se.

The main contribution of the paper is the Foreebank data set itself[2] but we also carry out preliminary parsing experiments evaluating the accuracy of a PCFG-LA parser on Foreebank, examining the effect of grammatical errors on parsing and experimenting with different training models.

## 2 Related Work

Other treebanks of English web text include the English Web Treebank (aka Google Web Treebank) (Mott et al., 2012), the small treebank of tweets and football discussion forum posts described in Foster et al. (2011) and the tweet dependency bank described in Kong et al. (2014). The English Web Treebank is a corpus of over 250K words, selected from blogs, newsgroups, emails, local business reviews and Yahoo! answers. It adapts the Penn Treebank (Marcus et al., 1994) and Switchboard (Taylor, 1996) annotation guidelines to address the phenomena specific

[1]http://community.norton.com

[2]www.computing.dcu.ie/mt/confidentmt.html

to this type of text. The annotation of the 1000-sentence treebank described in Foster et al. (2011) is based on the Penn Treebank, whereas the annotation of the treebank described in Kong et al. (2014) is dependency-based. The *French Social Media Bank* developed by Seddah et al. (2012) is a treebank of 1,700 French sentences from various type of social media including Facebook, Twitter and discussion forums (video game and medical). An extended version of the FTB-UC annotation guidelines (Candito and Crabbé, 2009) is employed during annotation and subcorpora containing particularly noisy utterances are identified.

The main difference between Foreebank and other web/social media treebanks is that grammatical errors in the Foreebank sentences are marked and corrected as part of the annotation process. Error annotation not only provides more insight into this type of text but it also enables us to directly measure the effect of these errors on parsing accuracy and leaves open the possibility of performing joint parsing and error detection by directly learning the error annotation during parser training.

A learner corpus (Granger, 2008) contains utterances produced by language learners and serves as a useful resource for researchers in second language acquisition, computational linguistics and computer-aided language learning. We can also compare Foreebank to a learner corpus since both contain utterances that are potentially ungrammatical and because in a learner corpus the errors are often annotated, as they are in Foreebank. Examples of learner corpora include the International Corpus of Learner English (Granger, 1993), the Cambridge Learner Corpus (Nicholls, 1999; Yannakoudakis et al., 2011), the NUS Corpus of Learner English (Dahlmeier et al., 2013) and the German Falko corpus (Lüdeling, 2008; Rehbein et al., 2012). In the last five years, there have been several shared tasks in grammatical error correction including the Helping Our Own (HOO) shared tasks of 2011 and 2012 (Dale and Kilgariff, 2011; Dale et al., 2012), and the CoNLL 2013 and 2014 shared tasks (Ng et al., 2013; Ng et al., 2014). With the exception of HOO 2011, all shared tasks involve error-annotated sentences from learner corpora. The annotation schemes vary from corpus to corpus but most involve marking the span of an error, classifying the error according to some taxonomy designed with L2 utterances in mind, and sometimes providing the correction or "target hypothesis" (Hirschmann et al., 2007).

As regards syntactic annotation of learner data, Dickinson and Ragheb (2009) propose a dependency annotation scheme for learner corpus based on the CHILDES annotation scheme (Sagae et al., 2007) developed for first language learners. They assume the developing language of learners as an interlanguage, as suggested by Díaz-Negrillo et al. (2010), and annotate it as is. They use two POS tags and two dependency labels for error cases: one for the surface form and one for the intended form. Rosén and De Smedt (2010) criticise the approach of Dickinson and Ragheb (2009) involving "annotating language text as is" arguing that interpretation of the language is required at all annotation levels. They use NorGram, a Lexical-Functional Grammar for Norwegian, to annotate a learner corpus with constituency structure, functional structure and semantic structure, with the purpose of providing a means to search for the syntactic context in which learner errors occur. Nagata et al. (2011) describe an English learner corpus which has been manually annotated with POS tags and shallow syntax, introducing two new POS tags and two new chunk labels for errors.

## 3 Building the Foreebank

The Foreebank treebank contains 1000 English sentences and 1000 French sentences. The English sentences come from the Symantec Norton technical support user forum. Half of the French sentences come from the French Norton forum and the other half are human translations of sentences from the English forum. Four annotators were involved in the annotation process. Their main task was to correct automatically parsed phrase structure trees using an annotation tool developed for this project.[3] The English annotators were guided by the Penn Treebank bracketing guidelines and a Foreebank-adapted version of the English Web Treebank bracketing guidelines. The French annotators used the French treebank (FTB) (Abeillé et al., 2003) guidelines, following the SPMRL strategy for multiword expressions (Seddah et al., 2013; Candito and Crabbé, 2009). The two primary annotators, one for French and one for English, annotated all the data for their

[3]The Stanford parser (Klein and Manning, 2003) was used to parse the English data and the Berkeley parser (Petrov et al., 2006) was used for the French sentences.

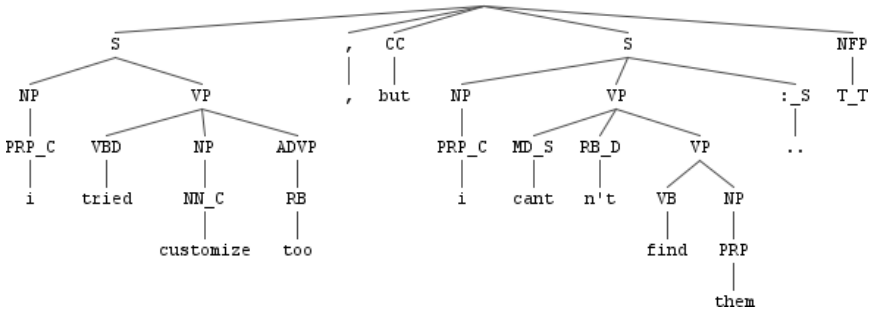


Figure 1: The Foreebank annotation of *i tried customize too , but i cant find them .. T_T* corrected as *I tried Customize too , but I ca n't find them ... T_T*

| Suffix | Explanation | Example | | $FB_{en}$ | | $FB_{fr}$ | |
|---|---|---|---|---|---|---|---|
| | | English | French | # | % | # | % |
| _D | Deleted token | It fixed [*the*] problem. | Cela a résolu [*le*] problème. | 170 | 1.10 | 56 | 0.29 |
| _X | Extraneous tokens | It fixed *the* my problem. | Cela a résolu *le* mon problème. | 35 | 0.23 | 17 | 0.09 |
| _W | Wrong form error | It *fix* the problem. | Cela *résoudre* le problème. | 69 | 0.45 | 43 | 0.22 |
| _S | Misspelled token | It fixed my *prbolem*. | Cela a résolu mon *prbolème*. | 81 | 0.53 | 117 | 0.60 |
| _C | Capitalisation error | *it* fixed my problem. | *cela* a résolu mon problème. | 161 | 1.00 | 194 | 1.00 |
| _B | Broken token | It *fix ed* my problem. | Cela a *réso lu* mon problème. | 2 | 0.01 | 12 | 0.06 |
| _I | Innovative initialism | I have problem *w/* this software. | J'ai un problème *av.* ce logiciel. | 1 | 0.01 | 7 | 0.04 |
| _M | Merged sentences | It fixed the problem Thank you. | Cela a résolu le problème Merci. | 3 | 0.30 | 29 | 2.90 |

Table 1: Foreebank Error Suffixes. The last two columns refer to their frequency.

language. The two secondary annotators annotated a 100-sentence subset. Inter-annotator agreement was calculated by measuring the Parseval F1 of trees produced by the secondary annotators against those produced by the primary annotators. For English this was 88 and for French it was 86.7.

Prior to correcting a parse tree produced by the automatic parser, the annotators are asked to correct any errors they find in the sentence.[4] The corrected text is entered in a field of the annotation tool. As part of the syntactic annotation process, errors are marked by appending an error suffix to the preterminals of the affected words in the tree. The error suffixes used in Foreebank are listed in Table 1 and an example tree from Foreebank is shown in Figure 1. There are three kinds of substitution error suffixes: `C` for marking problems with capitalisation, `S` for marking spelling errors and `W` for marking the wrong form of a word which encompasses inflection errors (*they* instead of *them*), real-word spelling errors (*test* instead of *text*) and lexical choice errors (*desk* instead of *chair*). The POS tag of the corrected form is used in the tree instead of the POS tag of the incorrect form.[5] Although this annotation scheme contains fewer error types than the taxonomies used for learner corpora, its granularity increases when the error suffixes are interpreted in the syntactic context in which they occur. For example, we can distinguish a missing determiner (`DT_D`) from a missing preposition (`IN_D`).

The "sentences" that the annotators see are the result of passing the forum text through an automatic sentence splitter (NLTK[6]) and tokeniser (in-house). This is another important difference between Foreebank and the English Web Treebank (EWT). In the EWT, sentence boundary detection and tokenisation has been carried out manually before annotation. Both approaches are valid but ours was chosen in order to stay closer to the more realistic scenario of less than perfect automatic preprocessing tools. This means that annotators have a special class of errors that result from noisy sentence splitting and tokenisation that must be marked during annotation.

There are two types of sentence splitting errors: merged sentences such as (1) in which a sentence boundary was not detected before the word *When* due to the use of a comma instead of a full stop, and split sentences such as (2).

[4] Minimal correction is encouraged to prevent annotators from rewriting the sentence in their preferred writing style. Instead they are instructed to just focus on fixing the errors.

[5] An alternative would have been to use one POS tag for the erroneous form and one for the corrected form, either combined a la Nagata et al. (2011) or separate a la Dickinson and Ragheb (2009).

[6] `http://www.nltk.org/`

(1) `7. Combofix will start, **When** it is scanning don't move the mouse cursor inside the box,`

(2) `The questions to <CompanyName>:`

Merged sentences are gathered under one root node with the error suffix `M` (e.g. `S_M`) , and split sentences are annotated as if they are standalone.

Tokenisation problems can also be categorised as merged (3) or split (4 and 5). Merged tokens are treated as a combination of a spelling error (*whenI* instead of *when*) and a deleted token (*I*). When the split is morphological as in (4), they are tagged with the POS tag of the whole intended token, along with the error suffix `B` (for "broken"). So in (4), the POS tag of *anti* would be annotated as `NN_B` and the POS tag of *vir* as `NN_B`. When there is no such clean morphological break (as in (5)), the first token is treated as a spelling error and the second as an extraneous token.

(3) `**whenI** tried to use ...`

(4) `he should buy **anti vir** programs`

(5) `**i t** keeps causing <ProductName> to lock up ...`

## 4 Analysing the Foreebank

Table 2 presents the average and the maximum sentence length in Foreebank, and, for comparison, WSJ and FTB. It also gives the out-of-vocabulary (OOV) rate of these data sets with respect to the WSJ and FTB. The Foreebank sentences are shorter on average than the WSJ and FTB sentences. The table also shows that the OOV rate of Foreebank with respect to WSJ/FTB is high: 33.3% for English and 39.1% for French. These numbers can be compared to the OOV rate of the WSJ test set with respect to its training set which is 13.2% and the FTB which is 20.6%. The higher OOV rate for the French Foreebank compared to the English is most likely due to the larger size of the WSJ compared to the FTB. The OOV rate of the English Foreebank is more than 2.5 times as large as that of the WSJ test set, while the OOV rate of the French Foreebank is less than 2 times as large as that of the FTB test set. This suggests that a bigger performance drop due to unknown words should be expected in parsing the English Foreebank sentences than the French.

The last four columns in Table 1 display the absolute and relative frequency of each error suffix. In sum, it seems that capitalisation is the major error type in Foreebank especially in the French data. Deleted tokens are also a major source of problem on the English side. Most of the capitalisation errors involve proper nouns (e.g. product names) and most of the deleted tokens are cases of missing punctuation. Overall, the errors occur on only a small fraction of the tokens in both data sets. We also calculate the edit distance between each Foreebank sentence and its correction by summing the number of error suffixes and dividing by the maximum of the original and corrected sentence lengths. The average edit distance for the English section of Foreebank is 0.04 and for the French section is 0.03. Despite the existence of some near-to-incomprehensible sentences, the overall error level is very low.

| | FB$_{en}$ | WSJ | FB$_{fr}$ | FTB |
|---|---|---|---|---|
| Avg. sentence length | 15.4 | 23.8 | 19.6 | 28.4 |
| Max. sentence length | 89 | 141 | 86 | 260 |
| OOV rate (%) | 31.6 | - | 33.6 | - |

Table 2: Characteristics of the English (FB$_{en}$) and French (FB$_{fr}$) Foreebank compared with those of the WSJ and FTB. The OOV rates are computed with respect to WSJ and FTB.

| **English** | FB$_{en}$ All | WSJ Test | **French** | FB$_{fr}$ All | FTB Test |
|---|---|---|---|---|---|
| WSJ$_{all}$ | 77.0 | - | FTB$_{all}$ | 76.3 | - |
| WSJ$_{train}$ | 75.4 | 89.6 | FTB$_{train}$ | 76.0 | 81.3 |

Table 3: Foreebank and WSJ/FTB test set results.

## 5 Parsing the Foreebank

We first evaluate newswire-trained parsers on Foreebank, using our in-house PCFG-LA parser with the max-rule parsing algorithm (Petrov and Klein, 2007) and 6 split-merge cycles. The English model is trained on the entire WSJ and the French model on the entire FTB. For comparison, we parse the WSJ/FTB and so we additionally use models trained only on the training sections. We remove the error suffixes and any D-suffixed nodes (representing deleted words) from the gold Foreebank trees before evaluation. The results are shown in Table 3. As expected, we see a significant drop for both languages when we move from in-domain data to Foreebank. Compared to parsing the English side of Foreebank, the performance drop for French is relatively smaller: the former drops 14.2 points from 89.6 $F_1$ points to

75.4 and the latter 5.3 points from 81.3 to 76. This suggests that, either the French parsing model is better generalisable to the forum text, or alternatively, that the FTB test set is more distant from its training set than the WSJ one. The second hypothesis is more likely because 1) it is on par with the OOV rate observed in Section 4, and 2) the performance of the English and French parsers are close on Foreebank but further apart on the newswire test sets. The effect of using the entire WSJ and FTB instead of only their training sections is also worth noting. While adding the WSJ development and test sets (about 5,500 sentences, a 14% increase) improves the $F_1$ of English parsing by 1.6 points, the 2,500 FTB development and test sentences (a 25% increase) have little effect on the French parsing, suggesting that either these new sentences are still not enough or do not bring additional information to the parsing model.

Since the annotators correct the errors made by the forum users, we are able to parse the corrected versions of the Foreebank sentences and examine how accurately they are parsed compared to the original sentences. We use the WSJ$_{all}$ and FTB$_{all}$ parsing models described above. Correcting the user errors before parsing leads to an improved parsing $F_1$ of 78.6 for the English sentences, an increase of 1.6 points (2%). On the other hand, a smaller impact is observed on the French sentences where the edited sentences receive an $F_1$ of 77.1 (an increase of 0.8 points). Referring to the distribution of error suffixes in Table 1, this suggests that the inserted and deleted tokens may have a larger effect on parser error than the substituted tokens, as their number is higher for English. This can be explained by considering that many substitution errors are capitalisation errors, typically involving a confusion between proper and common nouns, which tends not to affect the surrounding tree (Foster et al., 2011).

The simplest method to improve the accuracy of parsing Foreebank is to use it as supplementary training data. We do this using a 5-fold cross validation, in which Foreebank is randomly split into five parts, with each part used for the evaluation of the parsers trained on WSJ/FTB plus the other four parts. The results are shown in Table 4. Combining the larger treebank and Foreebank improves the $F_1$ by 2.6 points for English and 3.2 for French. Considering that Foreebank is orders of magnitude smaller than the WSJ/FTB, these gains

| English | | French | |
|---|---|---|---|
| Training Set | $F_1$ | Training Set | $F_1$ |
| WSJ$_{all}$ | 77.0 | FTB$_{all}$ | 76.3 |
| WSJ$_{all}$+FB$_{en}$ | 79.6 | FTB$_{all}$+FB$_{fr}$ | **79.5** |
| WSJ$_{all}$+5FB$_{en}$ | 80.1 | FTB$_{all}$+5FB$_{fr}$ | **79.5** |
| EWT | 75.0 | - | - |
| EWT+FB$_{en}$ | 79.0 | - | - |
| WSJ$_{all}$+FB$_{en}$+EWT | **80.3** | - | - |
| FB$_{en}$ | 71.1 | FB$_{fr}$ | 72.4 |
| FB$_{en_suf}$ | 70.2 | FB$_{fr_suf}$ | 71.8 |

Table 4: Training on Foreebank/WSJ/EWT/FTB and testing on Foreebank

are encouraging. We try to overcome the small size of Foreebank by 1) using the EWT as training data, and 2) increasing the weight of Foreebank by training on multiple copies of it. The EWT is not a substitute for the WSJ but it does provide a modest improvement when used in conjunction with Foreebank and WSJ. The replication of Foreebank trees has mixed results, providing a 0.5 point improvement for English and none for French.

In all experiments up to now, we have excluded the error suffixes from the Foreebank trees (during training and testing). We next try to directly learn trees containing the error suffixes (except for deleted tokens). That is, we use the original Foreebank trees containing the error suffixes for training and evaluate against Foreebank trees containing the error suffixes. The second last row of Table 4 shows the 5-fold CV results when the version of Foreebank without the error suffixes is used for training and the last row the results when the error suffixes are included. Including the suffixes decreases the accuracy, most likely due to the increased data sparsity caused by the suffixed tags.

## 6 Conclusion

We have introduced a treebank of technical forum sentences for English and French, based on an annotation strategy adapted to suit user-generated text in a realistic NLP setting. By marking the errors on the trees, we studied their prevalence as well as their impact on parsing and found that despite their low frequency, they do negatively affect parser performance, while not being the most important factor. Our next steps include learning error suffixes during a prior tagging phase and experimenting with the French Social Media Bank.

## Acknowledgments

This research has been supported by the Irish Research Council Enterprise Partnership Scheme (EPSPG/2011/102) and the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Dublin City University. We thank the reviewers for their helpful comments.

## References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a Treebank for French. In *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 165–187. Kluwer Academic Publishers.

Marie Candito and Benoît Crabbé. 2009. Improving Generative Statistical Parsing with Semi-supervised Word Clustering. In *Proceedings of IWPT*, pages 138–141.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Robert Dale and Adam Kilgariff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, pages 242–249.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings oftThe 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 54–62.

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage pos annotation for effective learner corpora in sla and flt. In *Language Forum*, volume 36, pages 139–154.

Markus Dickinson and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, pages 59–70.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL*, pages 359–369.

Jennifer Foster, Joachim Wagner, and Josef Van Genabith. 2008. Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of ACL: Short Papers*, pages 221–224.

Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From News to Comment: Benchmarks and Resources for Parsing the Language of Web 2.0. In *Proceedings of IJCNLP*, pages 893–901.

Jennifer Foster. 2010. cba to check the spelling investigating parser performance on discussion forum posts. In *Proceedings of NAACL*, pages 381–384.

Sylviane Granger. 1993. International corpus of learner english. In J. Aarts, P. de Haan, and N.Oostdijk, editors, *English Language Corpora: Design, Analysis and Exploitation*, pages 57–71. Rodopi, Amsterdam.

Sylviane Granger. 2008. Learner corpora. In Anke Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 259–275. Berlin:Mouton de Gruyter.

Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics*.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL*, pages 423–430.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–10012.

Anke. Lüdeling. 2008. Mehrdeutigkeiten und kategorisierung: Probleme bei der annotation von lernerkorpora. In M. Walter and P. Grommes, editors, *Fortgeschrittene Lernervarietten*, pages 119–140. Niemeyer, Tbingen.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the 1994 ARPA Speech and Natural Language Workshop*, pages 114–119.

Justin Mott, Ann Bies, John Laury, and Colin Warner. 2012. Bracketing Webtext: An Addendum to Penn Treebank II Guidelines. Technical report, Linguistic Data Consortium.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of ACL-HLT*, pages 1210–1219.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

D. Nicholls. 1999. The cambridge learner corpus – error coding and analysis. In *Summer Workshop on Learner Corpora*, Tokyo, Japan.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Proceedings of HLT-NAACL*, pages 404–411.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, 59.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact and Interpretable Tree Annotation. In *Proceedings of COLING-ACL*.

Ines Rehbein, Hagen Hirschmann, Anke Lüdeling, and Marc Reznicek. 2012. Better tags give better trees or do they? *Linguistic Issues in Language Technology*, 7(10).

Victoria Rosén and Koenraad De Smedt. 2010. Syntactic annotation of learner corpora. *Systematisk, variert, men ikke tilfeldig*, pages 120–132.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. High-accuracy annotation and parsing of childes transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *Proceedings of COLING*, pages 2441–2458.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182.

Ann Taylor. 1996. Bracketing Switchboard: An Addendum to the Treebank II Guidelines. Technical report.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of ACL*, pages 180–189.